

Análisis de datos para la promoción de la salud

Henry Augusto Salazar Blanco

Julio 1, 2021

Prevención de la sarcopenia basada en datos.

La sarcopenia es un síndrome que se caracteriza por la pérdida generalizada y progresiva de fuerza y masa del músculo esquelético, con un riesgo de resultados desfavorables como discapacidad física, baja calidad de vida y muerte (Cruz-Jentoft, y otros, 2010). Sorprendentemente, el criterio clínico que podría ayudar a identificar las personas con discapacidad de la función muscular, es aún insuficiente (Lauretani, y otros, 2003). El Grupo de Trabajo Europeo en Sarcopenia y Personas Mayores (EWGSOP, por sus siglas en inglés), recomienda usar en conjunto la presencia en la disminución de masa muscular y reducción de la función muscular (fuerza o rendimiento) para el diagnóstico de sarcopenia. Por tanto, el criterio para el diagnóstico de sarcopenia, basada en la documentación de la EWGSOP, son los siguientes síntomas:

- Pérdida de masa muscular
- Disminución de la fuerza muscular
- Pérdida del desempeño físico

1. Objetivos.

El objetivo de esta actividad es asimilar y poner en práctica los conocimientos adquiridos desde principios de semestre, y en este módulo en particular, en cuanto a:

- Programación en R
- La obtención de valor a partir del análisis de los datos
- La evaluación de un modelo de análisis de datos
- Visualización de datos
- Hacer recomendaciones prácticas en el contexto particular de la salud pública

2. Enunciado.

El objetivo de esta práctica es analizar una serie de datos históricos de pacientes de sarcopenia para determinar si existen una serie de síntomas más allá de los ya conocidos y que puedan ayudar a predecir la enfermedad antes de que se diagnostique. No todas las personas desarrollan un grado de sarcopenia que les inhabilite, y de las que la desarrollan, no todas lo hacen a la vez. Se ha hecho un seguimiento detallado de los síntomas que presentaban miles de pacientes de sarcopenia. La población en cuestión es el país de Atlantis, un país insular situado en el océano Atlántico, a unas 1.500 millas náuticas al oeste de las Islas Canarias, y los datos han sido recopilados durante un espacio de 25 años.

3. Objetivo de la práctica.

Eres un data scientist a quien han encargado la importante tarea de determinar los principales síntomas de la sarcopenia en la población de Atlantis. Contáis con los registros de la doctora Ellen Ripley, que consisten en la anotación de los síntomas de la sarcopenia en 9430 pacientes diferentes.

3.1. Explica qué son las reglas de asociación y proporciona algunos ejemplos de aplicación concretos, tanto en el contexto médico como otro más general. Los ejemplos deben estar apropiadamente referenciados.

Las **reglas de asociación** es una técnica de minería de datos empleada para encontrar patrones frecuentes en un conjunto de datos. Una definición formal se basa en el problema de extraer reglas de asociación, entre conjuntos de ítems en una base de datos de transacciones de los clientes. Eventualmente, cada transacción consiste de *ítems* o atributos comprados por un cliente en una visita al comercio (Amat Rodrigo, 2018). Por tanto, el término *transacción* hace referencia a cada grupo de eventos que están asociados de alguna forma, por ejemplo:

- La cesta de compras en un supermercado
- Los libros que un cliente compra en una librería
- Las páginas web que visita un usuario

El algoritmo *ARM* (*Association Rules Mining*, por sus siglas en inglés) surge por el interés de encontrar un tipo de reglas (Agrawal, Imielinski, & Swami, 1993) que tienen:

- **Mínimo soporte transaccional** s - la unión de ítems en el consecuente y antecedente de la regla está presente en un mínimo de $s\%$ transacciones en la base de datos.
- **Mínima confianza** c - por lo menos $c\%$ de transacciones en la base de datos que satisface el antecedente de la regla, también satisface el consecuente de la regla.

Un ejemplo de las reglas de asociación, considerando el modelo de un supermercado para la definición del algoritmo por parte de *Rakesh Agrawal, Tomasz Imielinski y Arun Swami* (Agrawal, Imielinski, & Swami, 1993, pág. 207); es la mayoría de personas que compran pan, también compran leche. Por tanto, en lenguaje matemático sería:

$$F(\text{Pan}) \Rightarrow \text{Leche}$$

La parte izquierda de la función se llama *antecedente* y la parte derecha se denomina *consecuente* (Jiménez Martínez, 2018). El descubrimiento de las reglas de asociación permite desarrollar estrategias de mercadeo con fines de aumentar las ventas, aunque también ha tenido aplicación en campos como la medicina, meteorología, detección de fraudes bancarios, prevención de fallas en aeronaves, entre otras áreas.

En cuanto a los estudios de mercadeo, se da el caso de una tienda de electrodomésticos, en donde los productos que los clientes compran con mayor frecuencia pueden ser ubicados en sitios distantes, mientras que los productos que se venden con mínima frecuencia se ubican en sitios cercanos a la vista de los clientes. Por ejemplo, si los clientes compran computadores y se inclinan por la compra de software financiero, entonces la ubicación del hardware puede ayudar a incrementar las ventas del software (De Moya Amaris & Rodríguez Rodríguez, 2003).

En el campo de la medicina, las reglas de asociación se han usado para identificar interacciones entre un conjunto de proteínas de VIH-1, y un conjunto de proteínas humanas con alta confianza, dado que la identificación entre un huésped viral y una proteína-proteína, tiene un enfoque importante y útil en el desarrollo de nuevos fármacos, permitiendo elegir el blanco de este tipo de interacciones (Mukhopadhyay, Maulik, Bandyopadhyay, & Eils, 2010).

El área de las imágenes médicas es adecuada para la extracción de información oculta o aún no descubierta, por medio del algoritmo de reglas de asociación. Imágenes de radiología (rayos X), así como imágenes digitales tales como Resonancia Magnética (MR, Magnetic Resonance), Ultrasonografía (US), Radiografía Computada Proyectada (CR, Projectional Computed Radiography), Tomografía Computada (CT, Computed Tomography) y Tomografía de Emisión de Posición (PET, Position Emission Tomography) se han usado ampliamente para tomar decisiones en cuanto al diagnóstico y pronóstico de una enfermedad. La naturaleza compleja de las imágenes médicas indica que al representar completamente dichas imágenes por medio de *ARM*, cada imagen o *transacción* tiene un amplio número de atributos o *ítems* (Olukunle & Ehikioya, 2002).

3.2. Caracterización del dataset.

Con el fin de realizar de manera sencilla el tratamiento de los datos, se ha migrado el conjunto de datos a un formato CSV, sin alterar los registros originales del archivo.

Para realizar el proceso de Extracción, Transformación y Carga de los datos, esencial para analizar la serie de los datos históricos; se requiere el cargue de unas librerías definidas por el lenguaje de programación R.

```
library(readr)
library(plyr)
library(dplyr)
library(arules)
library(arulesViz)
library(tidyverse)
library(knitr)
library(ggplot2)
library(lubridate)
library(tidyr)
library(sqldf)
```

```
#Se carga el conjunto de datos
sarco <- read.csv("sarcopenia_data.csv")
#Despliega las cabeceras o nombres de las columnas,
#las cuales corresponden a las variables.
names(sarco)
```

Exploración y pre-procesamiento de los datos

```
## [1] "PatientID" "Symptom"
```

```
#Permite visualizar el total de registros.
nrow(sarco)
```

```
## [1] 20796
```

```
#Valida el total de variables o columnas del conjunto de datos.
ncol(sarco)
```

```
## [1] 2
```

```
#Ofrece una previsualización del tipo de dato de cada variable
#y su contenido, así como el número de registros.
str(sarco)
```

```
## 'data.frame': 20796 obs. of 2 variables:
## $ PatientID: int 1 2 2 3 3 3 4 5 5 5 ...
## $ Symptom : chr "Obesity" "Pancreatic_Neoplasms" "Pancreatic_Neoplasms" "Abdominal_Obesity" ...
```

```
#Selecciona los registros o filas completas
sarco <- sarco[complete.cases(sarco), ]
#Confirma el número de registros
nrow(sarco)
```

```
## [1] 20796
```

```
#Sentencia que limpia las filas que tienen valores vacíos,
#y guarda los nuevos registros en la variable sarco1
sarco1 <- filter(sarco, !(is.na(Symptom) | sarco$Symptom == ""))
```

```
#Confirma el nuevo número de registros, almacenados en sarco1
nrow(sarco1)
```

```
## [1] 20609
```

```
#Imprime los primeros cinco registros del conjunto de datos sarco1
head(sarco1, n=5)
```

```
## PatientID      Symptom
## 1         1      Obesity
## 2         2 Pancreatic_Neoplasms
## 3         2 Pancreatic_Neoplasms
## 4         3 Abdominal_Obesity
## 5         3      Pain
```

En el resultado de la sentencia anterior se visualiza que el paciente 2 (PatientID = 2) tiene duplicado el síntoma de Pancreatic_Neoplasms. Por tanto, para no disponer de datos duplicados y posteriormente generar información veraz, se procede a continuación a mantener sólo datos únicos.

```
#Sentencia SQL que permite mantener registros únicos por paciente.
#Es decir, elimina los valores duplicados como se visualizaba
#en la consulta anterior para sarco1.
sarco2 <- sqldf('SELECT DISTINCT * FROM sarco1')
#Confirma el nuevo número de registros,
#almacenados en sarco2,
#luego de eliminar valores duplicados en el conjunto de datos sarco1.
nrow(sarco2)
```

```
## [1] 18960
```

```
#Imprime los primeros cinco registros del conjunto de datos sarco2
head(sarco2, n=5)
```

```
## PatientID      Symptom
## 1         1      Obesity
## 2         2 Pancreatic_Neoplasms
## 3         3 Abdominal_Obesity
## 4         3      Pain
## 5         3 Chronic_Kidney_Failure
```

Eventualmente, el PatientID 2 no dispone ahora del valor duplicado, situación similar con los registros restantes, luego de eliminar 1649 valores duplicados.

```
#Imprime las frecuencias o cantidad de incidencias por síntoma y enfermedad.
sarco2 %>% count("Symptom")
```

```
## "Symptom"      n
## 1 Symptom 18960
```

```
#Crea el factor de Symptom para hacer el group_by. Es decir,
#se modifica la naturaleza de la variable Symptom con fines
#de hacer el conteo
sarco2 %>% mutate(Symptom = as.factor(Symptom))
```

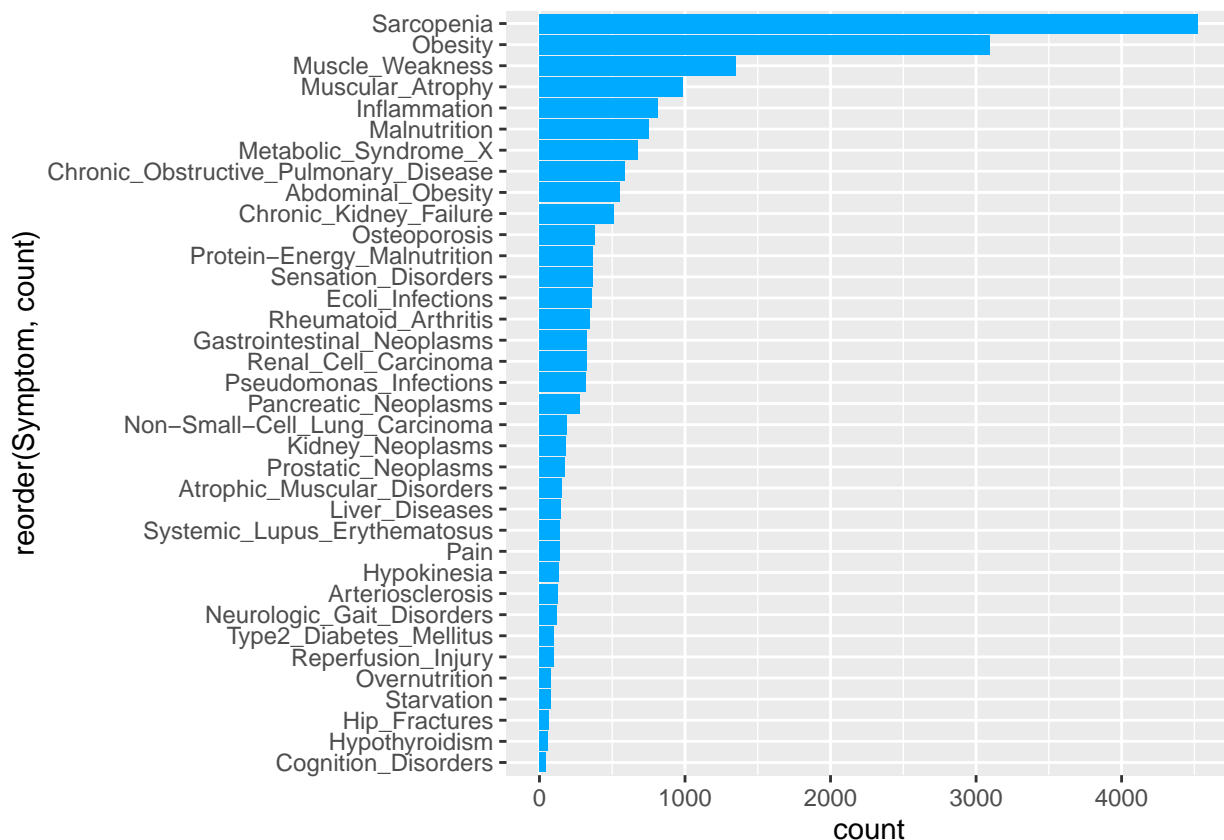
```
#Impresión del número de casos por síntoma y enfermedad.
detach("package:plyr", unload=TRUE)
sarco2 %>% group_by(Symptom) %>% summarize(count = n()) %>% arrange(desc(count))
```

```
## # A tibble: 36 x 2
```

```
## Symptom count
## <chr> <int>
## 1 Sarcopenia 4528
## 2 Obesity 3097
## 3 Muscle_Weakness 1350
## 4 Muscular_Atrophy 983
## 5 Inflammation 815
## 6 Malnutrition 753
## 7 Metabolic_Syndrome_X 680
## 8 Chronic_Obstructive_Pulmonary_Disease 585
## 9 Abdominal_Obesity 552
## 10 Chronic_Kidney_Failure 515
## # ... with 26 more rows
```

```
tmp <- sarco2 %>% group_by(Symptom) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

tmp %>% ggplot(aes(x=reorder(Symptom,count), y=count))+
  geom_bar(stat="identity",fill="#00abff")+
  coord_flip()
```



3.3. Estrategia optimizada para obtener reglas de asociación.

Antes de emplear cualquier algoritmo de regla de asociación, se requiere transformar los datos desde el conjunto de datos anterior, insertando transacciones de modo que todos los síntomas y enfermedades se encuentran en una sola fila.

```
#Realiza una ordenación del conjunto de datos,
#almacenándolo en la variable sarco2.sorted
sarco2.sorted <- sarco2[order(sarco2$Symptom),]
library(plyr)

#La función ddply() recibe el data frame, lo recorta en fragmentos basados
#en uno o más factores, para luego traer los resultados
#como un nuevo conjunto de datos llamado itemList.
itemList <- ddply(sarco2,c("PatientID"),function(df1)paste(df1$Symptom,collapse = ","))
```

Dado que las reglas de asociación requieren de las transacciones y los ítems, se procede a remover la columna de PatientID.

```
itemList$PatientID <- NULL
colnames(itemList) <- c("items")
```

Luego de generar el nuevo conjunto de datos, se crea un nuevo archivo en formato CSV (*sarcopenia_transaction.csv*) para examinar si las transacciones son correctas

```
#write.csv(itemList,"sarcopenia_transaction.csv", quote = FALSE, row.names = TRUE)
```

El nuevo conjunto de datos almacena las transacciones que se generan de acuerdo al diagnóstico médico.

```
#Carga y lee el conjunto de transacciones
tr <- read.transactions('sarcopenia_transaction.csv', format = 'basket', sep=',')
summary(tr)
```

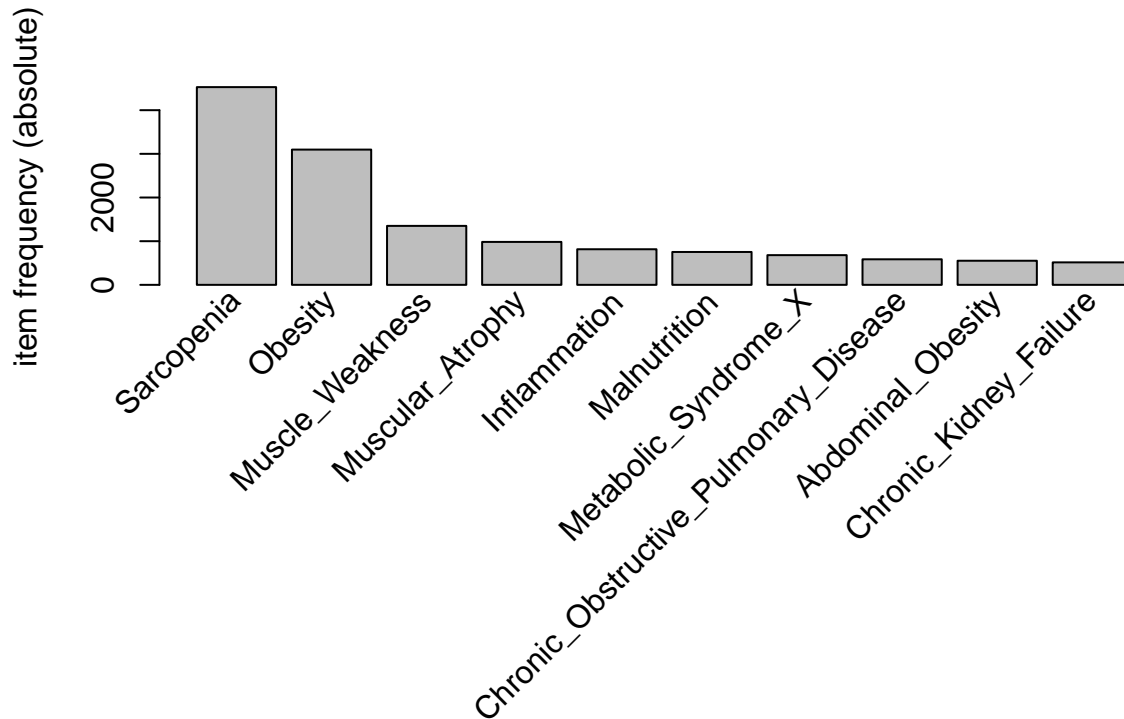
```
## transactions as itemMatrix in sparse format with
## 9410 rows (elements/itemsets/transactions) and
## 9446 columns (items) and a density of 0.0003191698
##
## most frequent items:
##      Sarcopenia      Obesity  Muscle_Weakness  Muscular_Atrophy
##      4528          3097          1350          983
##      Inflammation  (Other)
##      815          17597
##
## element (itemset/transaction) length distribution:
## sizes
##      1      2      3      4      5      6      7      8      9     10
##      1 3814 3092 1511  676  209   82   16    5    4
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000  2.000   3.000   3.015  4.000  10.000
##
## includes extended item information - examples:
##      labels
##      1      1
##      2     10
##      3    100
```

Según la última ejecución, se observa que existen 9410 transacciones, el cual corresponde al número de filas para el conjunto de datos reciente. Igualmente existen 9446 ítems, siendo los ítems el tipo de síntoma o enfermedad. De modo que las transacciones del presente ejercicio son los subconjuntos de los 9446 ítems.

Adicionalmente, la sentencia anterior permite hacer un breve análisis estadístico, dado la información útil que despliega el lenguaje de programación.

- Densidad: Es el número total de ítems que son padecidos por los enfermos, dividido por el the total de posibles ítems que están en la matriz o conjunto de datos. Se puede calcular cuantos ítems fueron padecidos, emleando la densidad como: $9410 \times 9446 \times 0.00032$.
- Observando las medidas de las transacciones, se evidencia que sólo una transacción corresponde a un ítem, 3814 transacciones para 2 ítems; siendo está última medida la transacción más grande. Esta medida indica que existe un gran número de enfermedades y síntomas padecidos por un pequeño número de ítems en cada transacción.

```
itemFrequencyPlot(tr, topN=10, type='absolute')
```



La figura anterior de ítems de frecuencias, está alineada con la figura 1, lo cual es acorde con la calidad de los datos, previa exploración y limpieza realizada al inicio del ejercicio.

3.4. Conjunto final de reglas

Para obtener las reglas de asociación se usa el algoritmo Apriori, disponible en la librería Arules desde el lenguaje de programación R para extraer los conjuntos de ítems frecuentes y las reglas de asociación. Se asigna un soporte de 0.001 y una confianza de 0.8 para generar todas las reglas que tienen por lo menos un soporte de 0.1% y una confianza mínima de 80%. Seguidamente se hace la ordenación de las reglas por confianza decreciente.

```
rules <- apriori(tr, parameter = list(supp=0.001, conf=0.8))
```

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##           0.8   0.1   1 none FALSE                TRUE     5   0.001   1
```

```

## maxlen target ext
##      10 rules TRUE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##      0.1 TRUE TRUE FALSE TRUE      2      TRUE
##
## Absolute minimum support count: 9
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[9446 item(s), 9410 transaction(s)] done [0.02s].
## sorting and recoding items ... [36 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [6 rule(s)] done [0.00s].
## creating S4 object ... done [0.00s].

rules <- sort(rules, by='confidence', decreasing = TRUE)
summary(rules)

## set of 6 rules
##
## rule length distribution (lhs + rhs):sizes
## 2 3 4
## 1 4 1
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##           2       3       3       3       3       4
##
## summary of quality measures:
##      support      confidence      coverage      lift
## Min.   :0.001063  Min.   :0.8095  Min.   :0.001169  Min.   :1.682
## 1st Qu.:0.001302  1st Qu.:0.8333  1st Qu.:0.001514  1st Qu.:1.732
## Median :0.001488  Median :0.8452  Median :0.001753  Median :1.757
## Mean   :0.002055  Mean   :0.8515  Mean   :0.002462  Mean   :1.770
## 3rd Qu.:0.001594  3rd Qu.:0.8643  3rd Qu.:0.001913  3rd Qu.:1.796
## Max.   :0.005420  Max.   :0.9091  Max.   :0.006695  Max.   :1.889
##      count
## Min.   :10.00
## 1st Qu.:12.25
## Median :14.00
## Mean   :19.33
## 3rd Qu.:15.00
## Max.   :51.00
##
## mining info:
## data ntransactions support confidence
##      tr          9410    0.001      0.8

```

El algoritmo Apriori después de la ejecución ofrece información útil:

- Se obtienen 6 reglas de asociación.
- Una longitud de 1 ítem tiene la mayoría de las reglas, según la distribución de reglas por longitud.
- El resumen de la calidad de medidas para el soporte, confianza, cobertura e incremento.
- La información sobre la extracción o minería de los datos.

Dado que el número total de reglas de asociación creadas por el algoritmo es de 6, se realiza la inspección de cada una de ellas.

```
inspect(rules[1:6])
```

	lhs	rhs	support	confidence	coverage	
## [1]	{Abdominal_Obesity, Chronic_Obstructive_Pulmonary_Disease, Malnutrition}	=> {Sarcopenia}	0.001062699	0.9090909	0.001168969	1.889
## [2]	{Inflammation, Pseudomonas_Infections}	=> {Sarcopenia}	0.001381509	0.8666667	0.001594049	1.801
## [3]	{Metabolic_Syndrome_X, Type2_Diabetes_Mellitus}	=> {Sarcopenia}	0.001275239	0.8571429	0.001487779	1.781
## [4]	{Metabolic_Syndrome_X, Reperfusion_Injury}	=> {Sarcopenia}	0.001594049	0.8333333	0.001912859	1.731
## [5]	{Gastrointestinal_Neoplasms, Malnutrition}	=> {Sarcopenia}	0.001594049	0.8333333	0.001912859	1.731
## [6]	{Hip_Fractures}	=> {Sarcopenia}	0.005419766	0.8095238	0.006695005	1.682

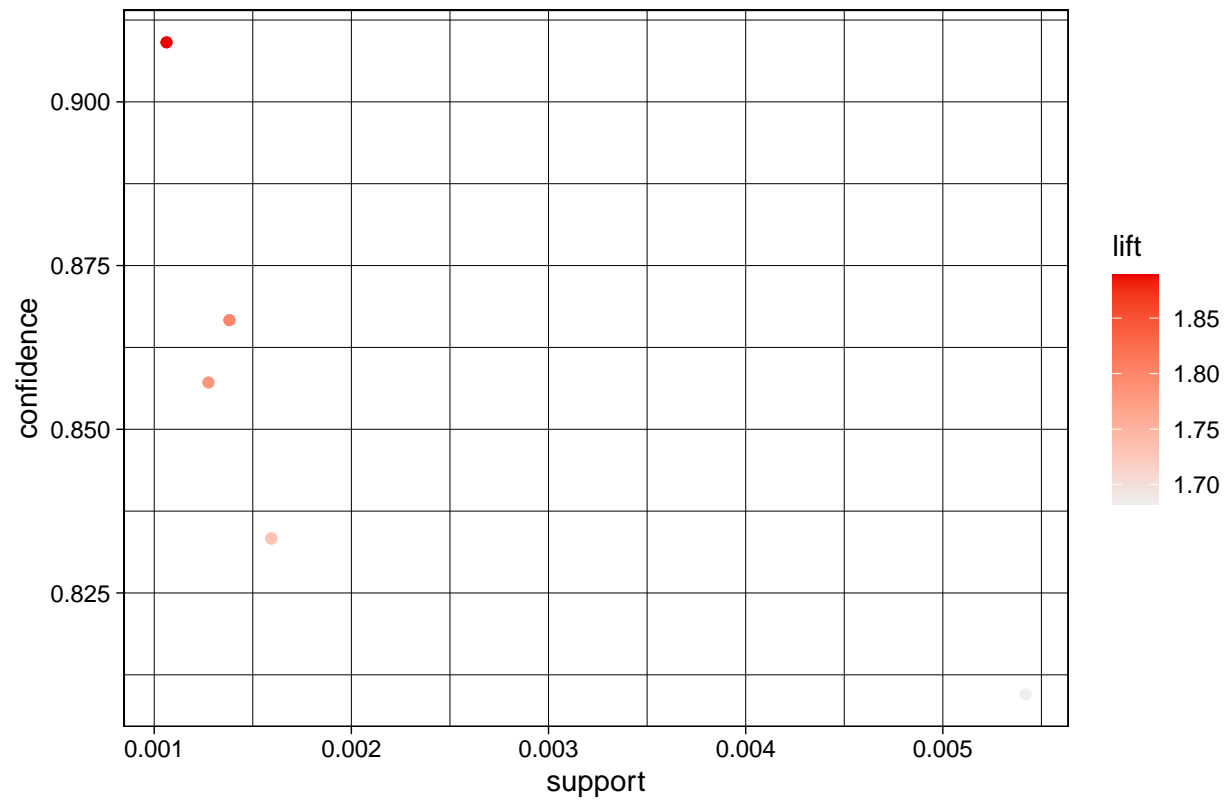
De acuerdo a la información generada, se tiene la siguiente interpretación: * El 91% de las personas que padecen obesidad, enfermedad crónica pulmonar y desnutrición también son propensas a padecer de sarcopenia. * El 86% de los adultos mayores enfermos de diabetes tipo 2 y síndrome metabólico X, tienen riesgo de enfermar por sarcopenia.

Gráficamente se puede hacer también un análisis a través de la siguiente sentencia:

Gráfica de forma dispersa

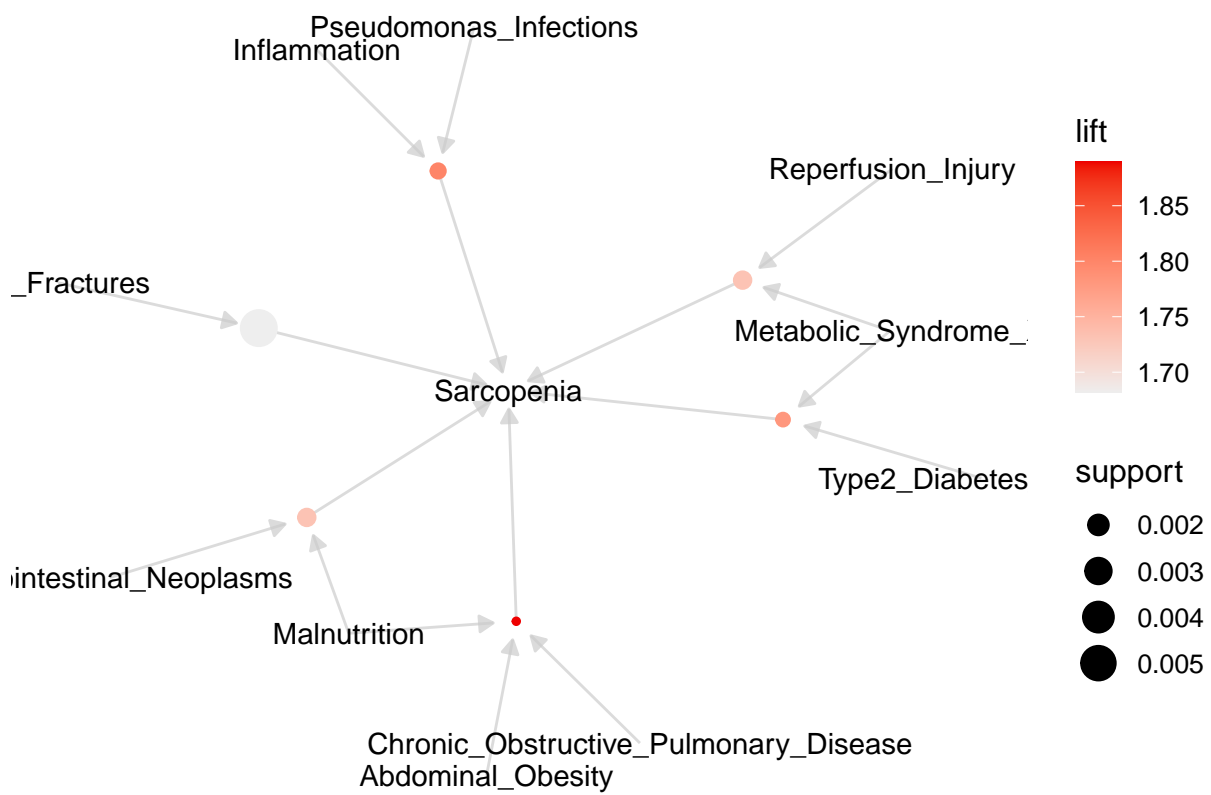
```
topRules <- rules[1:6]
plot(topRules, jitter=0, method = "scatterplot")
```

Scatter plot for 6 rules



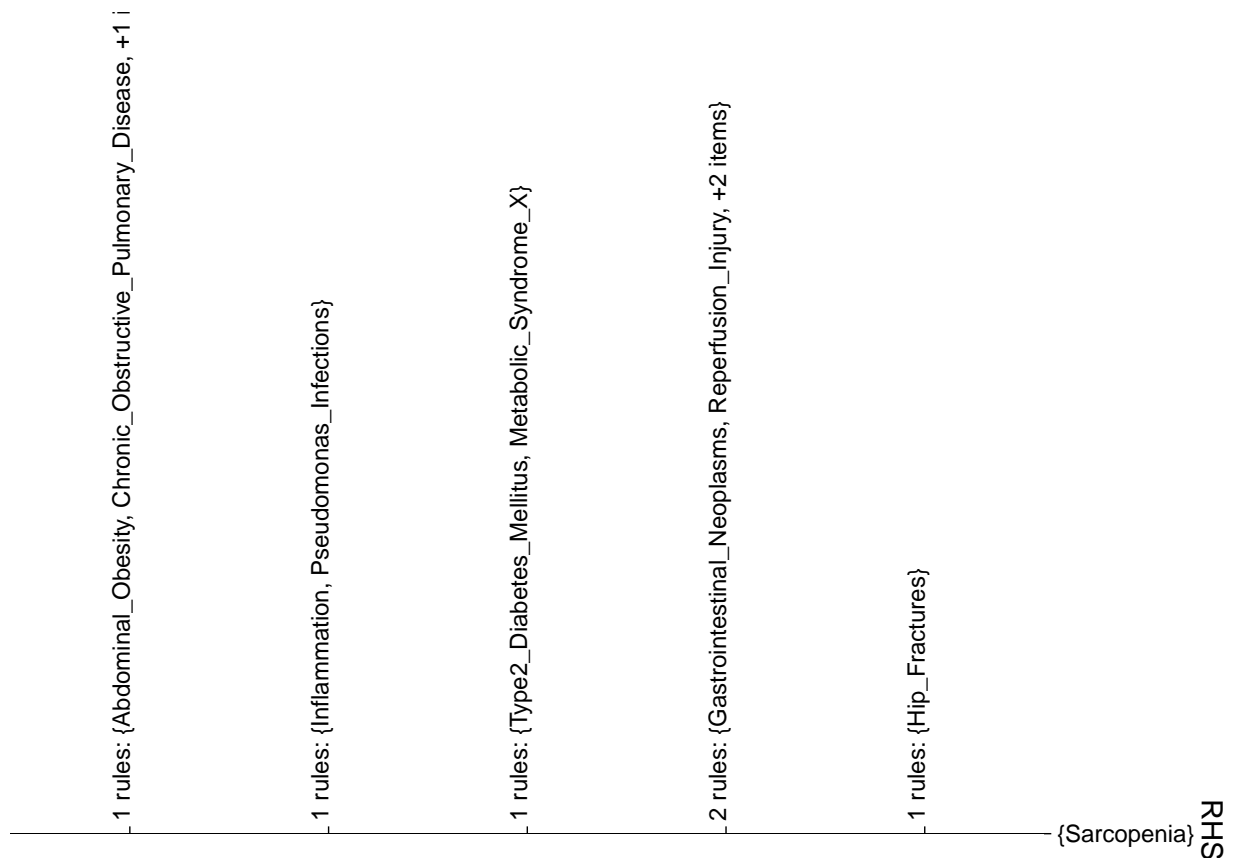
Gráfica en forma de grafo

```
plot(topRules, method="graph")
```



Gráfica de forma agrupada

```
plot(topRules, method = "grouped")
```



3.5. Estrategia basada en el dato para la gestión de la sarcopenia en Atlantis.

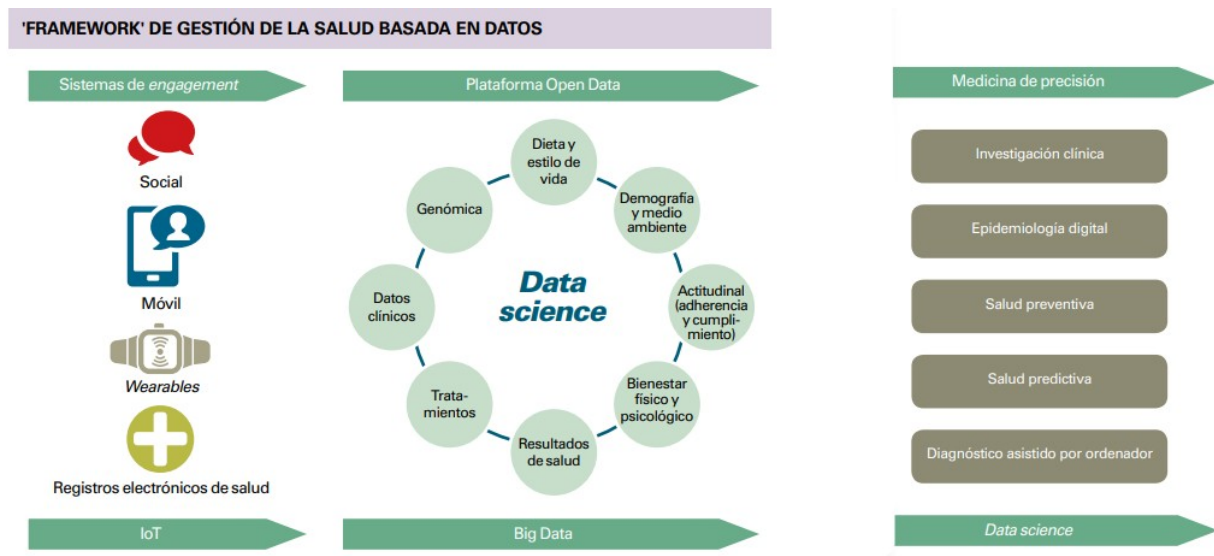
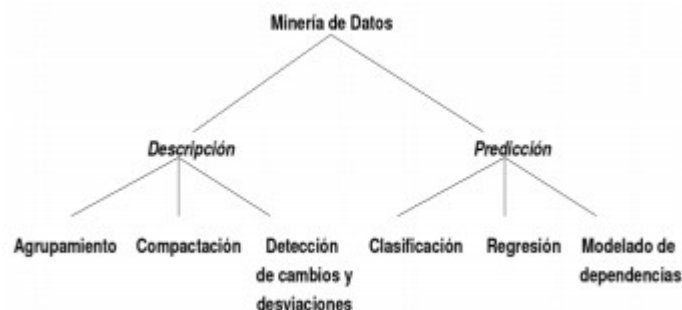


Gráfico tomado de “Hacia un cuidado de la salud basado en los datos”, Juan José Casado, Harvard Deusto Business Review

La estrategia de Gestión de la Sarcopenia en Atlantis debe procurar: mayor conocimiento de los pacientes, medicina personalizada, privacidad de información, superación de barreras culturales o legislativas, e integración de todos los actores de la salud (profesionales sanitarios, farmacéuticas, aseguradoras, sistemas de

provisión médica, pacientes y sociedad en su conjunto).(Casado, 2016). La cadena de gestión en el sistema sanitario de información inicia con el Engagement, es decir, con el involucramiento de los pacientes en el cuidado de su salud. En el caso de la sarcopenia con la población de adultos mayores y las variables asociadas a esta condición. (diabetes, hipertensión, tabaquismo, entre otras). En este nivel, resulta muy útil la incorporación de los mecanismos de IoT (Internet de las cosas) tales como Wearables, Apps, Telefonía móvil, Redes sociales Registros e historia clínica electrónica. Los wearables para la sarcopenia permitirían obtener información como: peso, índice de masa corporal, circunferencia del brazo, actividad física, consumo de proteína. Frente a esos datos de entrada, las reglas de asociación permitieron establecer patrones descriptivos y predictivos del envejecimiento en esta situación, y con base en eso, de desarrollaría una Plataforma Open Data de software libre, código abierto y acceso libre que favorezca el tratamiento permanente de los datos a gran escala que se generan en cada situación. La estrategia de salud, como proceso integral incluiría: Datos clínicos, Genómica, Dieta y Estilo de vida, Demografía y medio ambiente, Bienestar físico y psicológico, Intervenciones, Resultados de salud y Aspecto actitudinales como adherencia y cumplimiento de tratamientos. Dicha estrategia de gestión de la sarcopenia integrada a un sistema sanitario de información, sería un proceso cíclico y de 360°, que se retroalimenta permanentemente, pues la investigación científica y la minería de datos buscarían siempre afinar más las reglas de asociación, de tal forma que se obtengan mejores niveles de cobertura y confiabilidad. Acciones puntuales de esta estrategia serían (Casado, 2016): uso de tecnologías digitales para lograr un mayor conocimiento del paciente; aplicación de Big Data en el proceso de investigación clínica y farmacéutica; incorporación de la medicina preventiva basada en datos para favorecer el automanejo del paciente; gestión poblacional que maximice los resultados de salud, disminuyendo costes; aplicación de modelos predictivos para el control epidemiológico y la práctica clínica diaria; capacitación de la inteligencia artificial para dar soporte a los profesionales en el diagnóstico. Este modelo de gestión de la sarcopenia integrado al sistema sanitario de información partiría de la ciencia de datos a procesos como salud predictiva, salud preventiva, epidemiología digital y medicina de precisión. Dicha medicina de precisión, o medicina personalizada integra la salud desde un modelo preventivo para proteger el bienestar y reducir la probabilidad con la que aparece una enfermedad; todo ello desde un enfoque lo más exacto posible e integrando la genética, el medio ambiente y el estilo de vida, de forma específica para cada usuario del sistema de salud. Finalmente, el presente modelo contribuiría a tratamientos más eficaces y a vivir más tiempo de forma saludable. Actualmente, los obstáculos para el modelo como el propuesto no son tecnológicos o de acceso a información, sino bloqueos culturales (políticos, económicos, educativos, organizacionales) (Casado,2016) que es necesario superar para que la ciencia de datos gestione un futuro con mejor calidad de vida para todos.

3.6. Limitaciones que tienen las reglas de asociación como herramienta de data mining.



Clasificación y relación entre los diferentes algoritmos de minería de datos (Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo Alejandro Ballesteros Román, Daniel Sánchez-Guzmán and Ricardo García Salcedo, 2013)

Las Reglas de Asociación son un instrumento muy importante para encontrar relaciones entre variables, sobre todo cuando se está manejando una gran cantidad de datos; por ésta razón pueden ser de gran apoyo para el área de la salud, por ejemplo, en la epidemiología, permiten lograr que los datos cumplan criterios para ser leídos de forma completa, agrupada, y con óptimos niveles de confiabilidad y comparabilidad (como ocurre en la estimación del riesgo de una enfermedad) (Merchán, 2011). No obstante, algunas limitaciones de las reglas de asociación, es que, con la minería, se pueden obtener un número muy grande de reglas, y al ser tantas,

algunas pueden “equivalentes, obvias o no útiles.” (Merchán, 2011) Frente a esta dificultad, se requiere invertir un mayor esfuerzo en los procesos de interpretación, evaluación, y elección de las reglas más pertinentes. Una segunda limitación se asocia al manejo de variables categóricas/variables numéricas. Como las reglas funcionan con variables categóricas, solo queda discretizar las variables numéricas. Esta condición influye, por tanto, en que se dé lugar a información irrelevante. Emplear reglas de asociación implica un tiempo de cómputo amplio, que para ciertas metas puede ser completamente ineficaz. Evidentemente un trabajo de minería de datos debe lograr un balance entre costo y beneficio. Los análisis científicos de salud deben cumplir parámetros de cobertura (lo que la regla predice) y confianza (porcentaje de aplicabilidad de la regla) y ello no debe exceder los recursos destinados para este proceso. Las reglas de asociación si pueden tener un carácter predictivo, lo cual no excluye otros resultados como la descripción, la estimación o el agrupamiento. El énfasis de la predicción está en los registros que se organizan de acuerdo a un comportamiento futuro. En este contexto la predicción tiene un valor enorme para la toma de decisiones inteligentes, seguras y acertadas en contextos sanitarios, empresariales y de mercadeo, entre otros. La predicción lleva consigo que se puedan seguir realizando acciones de investigación, desarrollo e innovación. Finalmente, por mínimo que sea, siempre puede haber un error en una predicción. Y esto no sería un problema. ¡También es esperanzador! Si se puede predecir con certeza inequívoca sobre la salud y la enfermedad... ¿Dónde quedarían la libertad, la creatividad, la esperanza? Como afirmó Aldous Huxley “La medicina avanza tanto que pronto estaremos todos enfermos” ... y si la minería de datos conduce a dicho avance, está en las manos del científico de datos encontrar constantemente un nuevo concepto de salud.

Bibliografía

- Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, 215.
- Amat Rodrigo, J. (Junio de 2018). Ciencia de Datos. Obtenido de https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion#Introducci%C3%B3n
- Casado, Juan José (2016) Hacia un cuidado de la salud basado en los datos”.*Harvard Deusto Business Review*.
- Cruz-Jentoft, A. J., Baeyens, J. P., Bauer, J. M., Boirie, Y., Cederholm, T., Landi, F., . . . Zamboni, M. (July de 2010). Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People. (O. Academic, Ed.) *Age and Ageing*, 39(4), 412-423. doi:<https://doi.org/10.1093/ageing/afq034>
- De Moya Amaris, M. E., & Rodríguez Rodríguez, J. E. (2003). La contribución de las reglas de asociación a la minería de datos. *Tecnura*, 96. Obtenido de <https://revistas.udistrital.edu.co/index.php/Tecnura/article/view/6175>
- Jiménez Martínez, C. (6 de 3 de 2018). RStudio Pubs. Obtenido de https://rstudio-pubs-static.s3.amazonaws.com/367334_353f1bbf1b3543e180bb9210e711a73f.html#fnref4
- Lauretani, F., Russo, C. R., Bandinelli, S., Bartali, B., Cavazzini, C., Di Iorio, A., . . . Ferrucci, L. (01 de November de 2003). Age-associated changes in skeletal muscles and their effect on mobility: an operational diagnosis of sarcopenia. *Journal of Applied Physiology*, 95(5), 1851-1860. Obtenido de <https://journals.physiology.org/doi/full/10.1152/jappphysiol.00246.2003>
- Marchán, Edgar, & Salcedo, Juan, & Aza, Teresa, & Figuera, Lourdes, & Martínez de Pisón, Francisco, & Guillén, Pablo (2011). Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas. *Ciencia e Ingeniería*, 32(2),55-60
- Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Eils, R. (2010). Mining Association Rules from HIV-Human Protein Interactions. *Proceedings of 2010 International Conference on Systems in Medicine and Biology*, 344-348. doi:10.1109/ICSMB.2010.5735401
- Olukunle, A., & Ehikioya, S. (2002). A Fast Algorithm for Mining Association Rules in Medical Image Data. *Proceedings of the 2002 IEEE Canadian Conference on Electrical & Computer Engineering*, 1181.